

Module 2: OMAmer for sequence placement into HOGs

Why sequence placement ?

	Speed	Input	Proteomes
Orthology database	Instantaneous	Known identifiers	Only in the database
Placement into HOGs	Few minutes	Few sequences or whole proteome	Any proteome
Orthology inference	Hours	Whole proteomes	Any proteomes (one or several)

What is OMAmer?

- ❖ Fast sequence placement into existing HOGs from the OMA Browsers
- ❖ More accurate than closest sequence matching for subfamily placement!

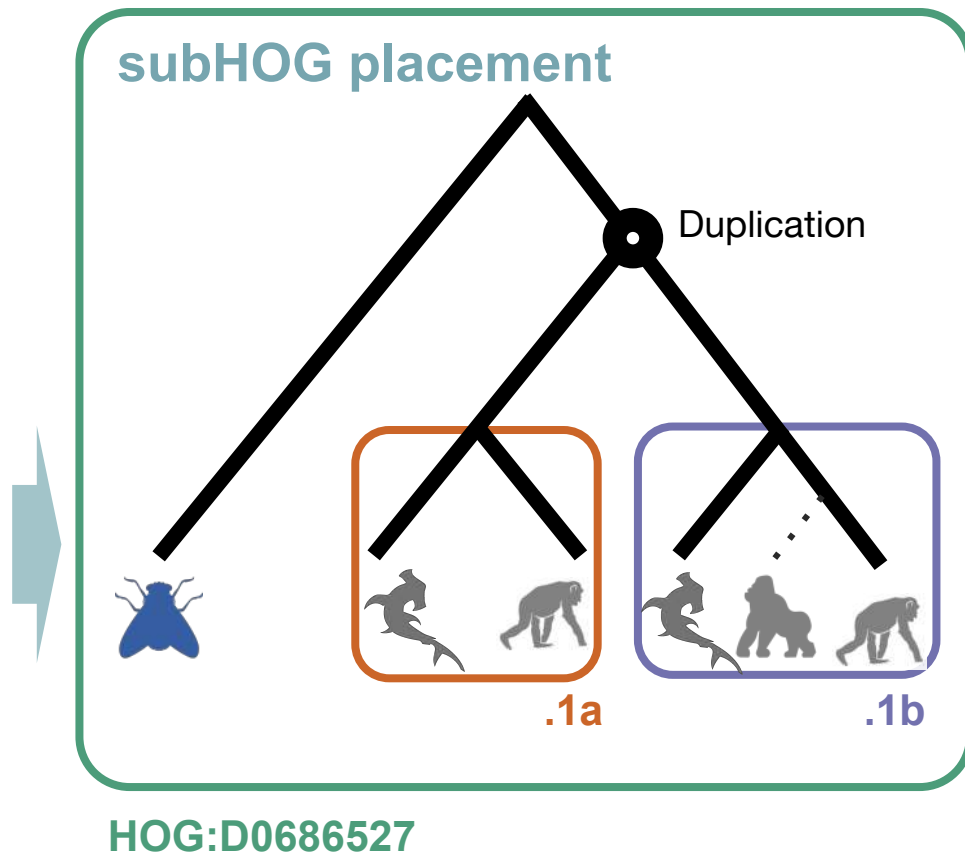
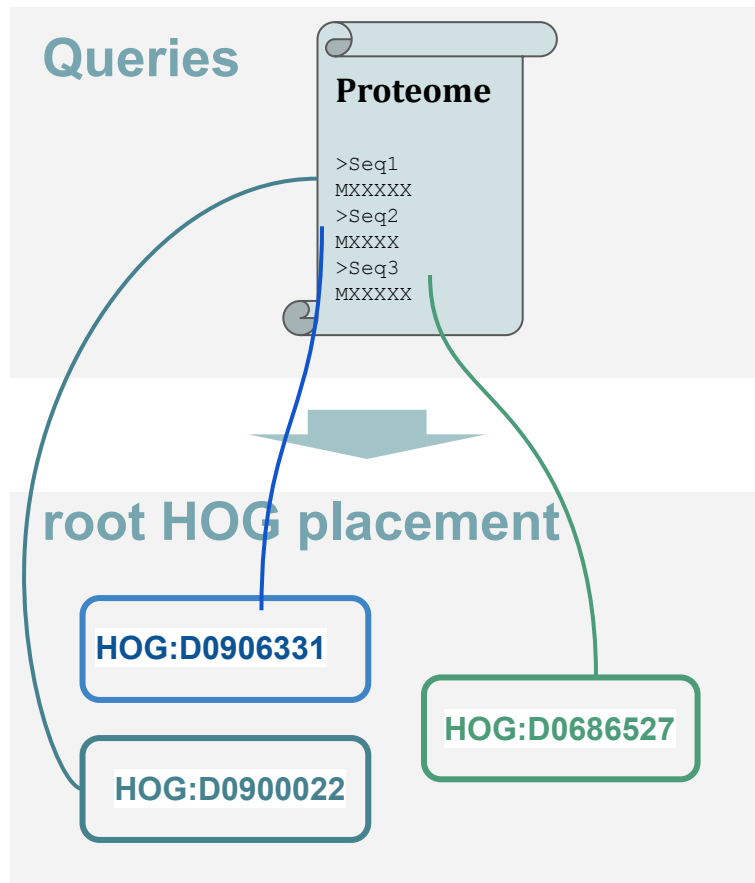
OMAmer: tree-driven and alignment-free protein assignment to subfamilies outperforms closest sequence approaches

Victor Rossier ^{1,2,3}, Alex Warwick Vesztröcy ^{1,2,3}, Marc Robinson-Rechavi ^{3,4,*}
and Christophe Dessimoz ^{1,2,3,5,6,*}



<https://github.com/DessimozLab/omamer>

OMAmer placement - principle



k-mer based placement

- ❖ **k-mers** : words of k characters in a sequences

Query sequence

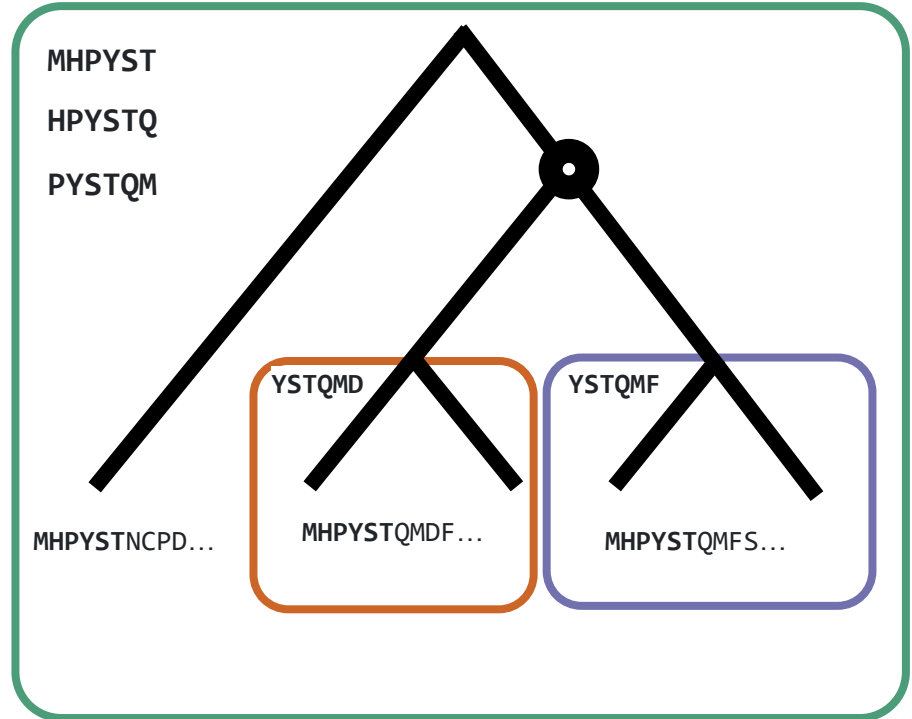
MHPYSTQMFS LQITVMEDSQ SDMSIELPLS

MHPYST
HPYSTQ
PYSTQM

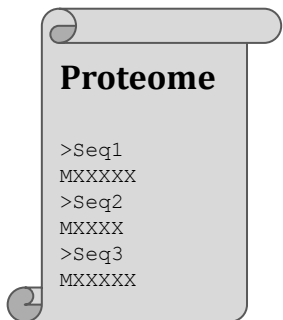
...
...
...

MSIELP
SIELPL
IELPLS

HOG



How to use OMAMer



Query sequences

FASTA format

From any species

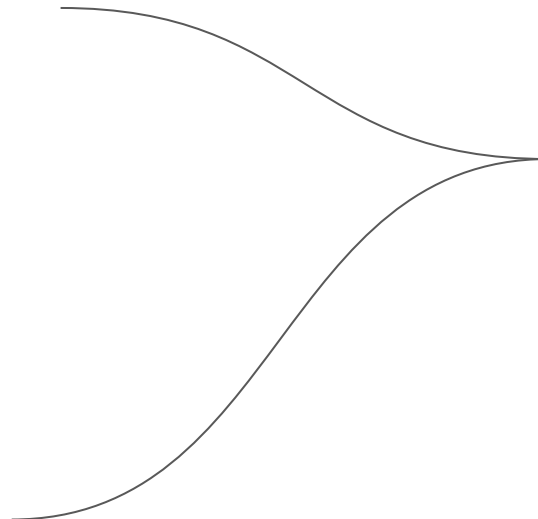


OMAMer database

HDF5 format

*Built with HOGs from the
OMA Browser*

```
omamer search --query query.fa --db db.h5 --output results.txt
```



```
Seq1 HOG:D0578800.1c.1d  
Seq2 HOG:D0571029  
Seq3 HOG:D0606120.3n
```

OMAMer output

Tab separated format

All HOG placements

Interpreting the output

qseqid	hogid	overlap	family-score	subfamily-score	qseqlen	subfamily-medianseqlen
Seq1	HOG:D0681219.3c	1.0	0.99701[...]	0.99701[...]	966	1065
Seq2	HOG:D0433152	1.0	0.99967[...]	0.99967[...]	303	334

❖ **qseqid** : Query identifier

Interpreting the output

qseqid	hogid	overlap	family-score	subfamily-score	qseqlen	subfamily-medianseqlen
Seq1	HOG:D0681219.3c	1.0	0.99701[...]	0.99701[...]	966	1065
Seq2	HOG:D0433152	1.0	0.99967[...]	0.99967[...]	303	334

- ❖ qseqid: Query identifier
- ❖ **hogid : Where the query is placed**

Interpreting the output

qseqid	hogid	overlap	family-score	subfamily-score	qseqlen	subfamily-medianseqlen
Seq1	HOG:D0681219.3c	1.0	0.99701[...]	0.99701[...]	966	1065
Seq2	HOG:D0433152	1.0	0.99967[...]	0.99967[...]	303	334

- ❖ qseqid: Query identifier
- ❖ hogid : Where the query is placed
- ❖ **Family score/subfamily score : Proportion of k-mer of the sequence in common with HOG (- Proportion expected by chance)**

Interpreting the output

qseqid	hogid	overlap	family-score	subfamily-score	qseqlen	subfamily-medianseqlen
Seq1	HOG:D0681219.3c	1.0	0.99701[...]	0.99701[...]	966	1065
Seq2	HOG:D0433152	1.0	0.99967[...]	0.99967[...]	303	334

- ❖ qseqid: Query identifier
- ❖ hogid : Where the query is placed
- ❖ Family score/subfamily score : Proportion of k-mer of the sequence in common with HOG (- Proportion expected by chance)
- ❖ **Overlap: distance between the first and last matching kmer on the sequence (0 to 1)**

Interpreting the output

qseqid	hogid	overlap	family-score	subfamily-score	qseqlen	subfamily-medianseqlen
Seq1	HOG:D0681219.3c	1.0	0.99701[...]	0.99701[...]	966	1065
Seq2	HOG:D0433152	1.0	0.99967[...]	0.99967[...]	303	334

- ❖ qseqid: Query identifier
- ❖ hogid : Where the query is placed
- ❖ Family score/subfamily score : Proportion of k-mer of the sequence in common with HOG
(- Proportion expected by chance)
- ❖ Overlap: distance between the first and last matching kmer on the sequence (0 to 1)
- ❖ **qseqlen and subfamily-medianseqlen : length of proteins in amino acids**

To remember

- ❖ Placement into HOGs allows to find gene families for **species not in the database**
- ❖ Can be used on **any number of sequences** - from one to whole proteomes
- ❖ Precise to the subfamily level but not a definitive proof of orthology
- ❖ Allows to still **take advantage of OMA Browser** wealth of data and features



OMA Academy

Welcome to the OMA Academy! Here, you will find online exercises which will aid you in becoming more familiar with orthology, phylogenies, and comparative genomics.

BACKGROUND

OMA ("Orthologous Matrix") is a method and database for the inference of orthologs among complete genomes. It can be found at omabrowser.org. Many of the exercises use the OMA browser as a starting point. The OMA pipeline can also run on custom genomic/transcriptomic data using the OMA stand-alone software, and it is even possible to combine precomputed data with custom data by exporting parts of the OMA database.

Tables of contents

1. [Exploring Orthology with the OMA Browser](#)
2. [OMAmer](#)
3. [FastOMA](#)
4. [Estimating a Species Tree](#)

Module 2: work until 12:10

Module 2: Fast placement of sequences into HOGs with OMAmer

Sometimes you might have a few protein sequences from a genome which is not in the OMA database and you want to quickly find out which genes they share homology with. Or perhaps you even want to do this with a whole proteome.

OMAmer is a command-line software that places a given protein sequence onto one of the gene families available in the input OMA database. In other words, OMAmer finds the most likely HOG where the input protein belongs. OMAmer is based on comparing *k*-mers (substring of the sequence of *k* length) between a query sequence and HOGs. Since it only searches for *k*-mers that are in common between sequences, it does not need a sequence alignment (which is usually computationally intensive) and is a very fast alternative to high-resolution homology determination when one is simply looking for the gene family a sequence belongs to.

[Back to home](#) / [Reset](#)

2.1 OMAmer setup and requirements	▼
2.2 Placing a few sequences into Hierarchical Orthologous Groups	▼
2.3 Placing a whole proteome	▼

```
source /workspace/conda/bin/activate
conda activate omacademy
```

```
nextflow FastOMA/FastOMA_light.nf --input_folder in_folder --output_folder out_folder
-resume
```

```
cd /workspace/SIBBiodiversityBioinformatics2023/Module3_FastOMA/expected_output/
```

Typo alert! (Module 3.1):

In this exercise, we will run FastOMA standalone to infer the orthology information for five yeast species. We already provided the proteomes of five species in the GitPod environment, located at

```
/workspace/SIBBiodiversityBioinformatics2023/Module3_FastOMA/working_dir/in_folder/proteome.
```

Click to go back, hold to see history

 .gitpod.yml	Changed path in the github file to reflect changing name of repo	yesterday
 LICENSE	Initial commit	5 days ago
 README.md	Update README.md	now

☰ README.md 

SIB Biodiversity Bioinformatics 2023

Teachers

- Natasha Glover
- Yannis Nevers
- Sina Majidian
- Christophe Dessimoz

FastOMA (temp)

FastOMA command line

```
cd /workspace/SIBBiodiversityBioinformatics2023/Module3_FastOMA/working_dir/  
nextflow FastOMA_light.nf --input_folder in_folder --output_folder out_folder
```



expected output structure for test data

Then, following files and folders should appear in the folder `out_folder` which was the argument.

```
$ls out_folder  
hogmap OrthologousGroupsFasta OrthologousGroups.tsv output_hog.orthoxml
```

